Predicting NFL Rushing Yards Using Bayesian Linear Regression

Matthew Kilgariff
mkilgar1@jhu.edu

Abstract

This paper will aim to explore how Bayesian statistics can be used to model success of a running back in the NFL. The dataset originates from Kaggle's *NFL Big Data Bowl 2020*, derived from AWS NextGen Stats. The paper will begin with a background of American Football, explanation and reasoning of the methodology, data exploration, and results. The data will be analyzed using Bayesian linear models, specifically Bayesian linear regression to predict number of rushing yards per attempt. The goal of the analysis is to construct probability densities that an NFL offensive coordinator or head coach can use to call plays intelligently. Five features were selected for the models: player weight, yard line, down, yards-to-go, and temperature. Results from an ordinary least squares fit were consistent with those of the Bayesian linear regression. The Bayesian approach allows for more situational information to be extracted from the data, such as probability of scoring or gaining a first down.

1 Purpose

1.1 Background of American Football

American football is one of the most complicated sports. Twenty two players, each with specialized roles, work together to move (or prevent the move of) the football into the endzone. The offense consists of four main position categories: the offensive line, who "block" defensive players from tackling the ball carrier; the receivers, who catch the ball on passing plays; the quarterback, who either passes or hands off the ball to a runner on each play; and the running back, who receives a handoff and runs with the ball towards the endzone. On each possession, an offense has four attempts, or downs, to move the ball ten yards. If successful, the downs reset, and the team gets four new attempts. This repeats until the ball crosses over the goal-line for a touchdown, a field goal is kicked, the ball is punted to the opposing team, or ball is turned over via a fumble, interception, or a failed fourth-down conversion.

On each running (also known as "rushing") play, the running back has to be agile to maneuver through the "holes" the offensive line creates with their blocks and avoid defenders attempting to tackle him, all while keeping possession of the football. The success on rushing plays are a crucial part of controlling field position, possession percentage, variety of play calling, and overall success of an NFL team. Downstream, this success affects win percentages, playoff success, and revenue for the team. How far a rusher can run the ball depends on his athletic prowess, the ability of his teammates to create rushing lanes, and the success (or failure) of the defense to disrupt the offense.

1.2 Why take a Bayesian approach?

NFL coaches spend nearly every hour of the day leading up their weekly match watching film, designing plays, and creating game plans. During games, they make tough decisions: whether to run

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

or pass, who gets and where they get the ball, and whether to attempt a fourth down conversion or kick or punt the ball. In this sense, play calls are extremely situational. NFL coaches tend to run the ball on early downs, when there are few yards-to-go, there is plenty of time remaining in the game, or there is either a favorable or a low-point differential.

Before the introduction of statistical methods in sports, NFL coaches used their intuition and "football knowledge" to estimate probabilities of success. If before a game, the coaches believe they can exploit the oppositions run defense but get stopped on every rush attempt, the coach may call less run plays. The probability, in his mind, of getting big runs has decreased as the game went on. Even though he may not know it, he is taking a Bayesian approach to calling plays.

An NFL analytics team could use a frequentist approach to predict rushing yards. However, this only gives a coach one number, as in, our prediction says our running back will get four yards on this play. This single number gives little information to the coach. He would rather have estimated probabilities of running five or more yards or getting a first down.

This paper will be investigating the how many yards a runner will gain or lose, as well as the probability of getting a first down on rushing attempts, using Bayesian linear regression and Bayesian classification techniques.

2 Literature Review

2.1 The Running Back

In a 2016 *New York Times* article, journalist Mike Tanier described the resurgence of the running back. Pass-to-rush ratios significantly decreased in the 2010s, as NFL coaches pushed to remove their one-note, pass-heavy stereotypes from the 1990s and 2000s. Arkes, however, in 2011 used logistic regression on first half NFL data to argue that controlling the passing game is the most important offensive key to success. Pelechrinis, in 2016, attempted to recreate ESPN's real-time win probability calculation. He showed that rushing attempts increased win percentage faster than passing attempts. This follows football intuition, as rushing plays are more likely to gain yards, and an unsuccessful gain keeps the clock running, unlike an incomplete pass. *Eldo.co* showed that rushing yards per attempt increased, passing attempts significantly increased, and leading rushers as a percentage of team rushing attempts increased over time. NFL teams are relying more on star running backs to make big plays, rather than using a committee of multiple runners.

2.2 Bayesian Analysis of NFL Data

Since the success of the "Moneyball" Oakland A's of the MLB, NFL teams have begun to use statistical methods to assist play calling and game plan construction. Each team uses *very* proprietary Bayesian methods to estimate first-down, win, and other probabilities. Coaches can use this analysis to compare estimated success of different plays given the situation on the field. These models are heavily protected, since a better model gives a team a competitive advantage. However, the league itself has published some of their research. Using the same data set as this paper, NFL.com constructed a Bayesian Network to create a quantitative metric of running back, team defensive, and offensive line skill. This analysis was then used to model a causal relationship between defensive formation and yards gained.

3 Proposed Method

3.1 Data Exploration

The data set used in this paper originates from Kaggle's *NFL Big Data Bowl 2020*. It contains data on every player's attributes, position, and velocity and acceleration, as well as game state information on every rushing play from the 2017, 2018, and 2019 regular seasons. In all, the data set has 31,007 plays, with 443 rushers from all 32 teams.

The dataset contains 49 features, one of which is the target yards gained, with roughly half considered "state variables": the weather, time, and location of the game and the performance of each team so far (i.e., points scored, yard line, yards-to-go). Other features include non-informative identifiers like

player ID and team names and where the player went to college. For the purposes of this paper, only quantitative data will be used. Categorical data like offensive or defensive formation will be ignored.

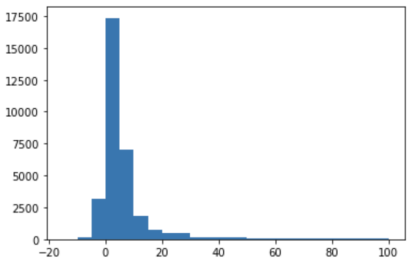


Figure 1: Distribution of yards Gained

The data is very uncorrelated (Figure 2), with the exceptions of player weight and height, temperature and week, and home and visitor score. This makes sense as a taller person will weigh more than a shorter. The NFL season begins in September and ends in January, corresponding with the season change, and teams score more points as the game goes on, so there are more points scored for both teams with little time remaining than a lot of time remaining.

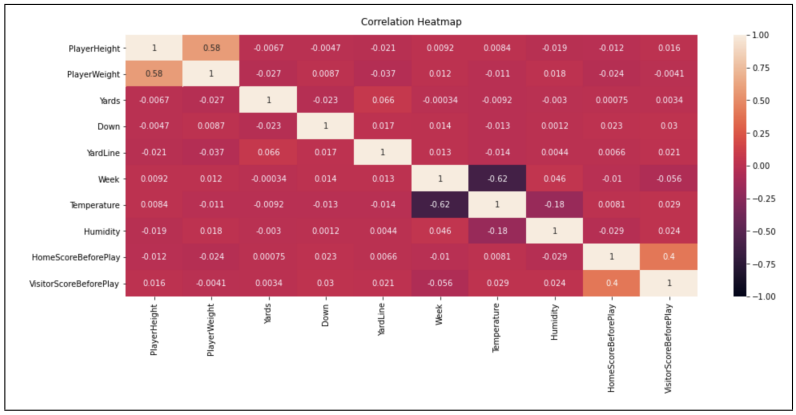


Figure 2: Factor Correlation Matrix

There are a few outliers. With player weight and height, there are examples of abnormally tall or heavy players. To remedy this, and to keep the focus of this paper on pure running backs. Defensive players, lineman, quarterbacks, and wide receivers will be removed. Big plays, those defined as runs for more than 25 yards, will also be excluded, as the magnitude of these plays depend on field position. If a running back can maneuver through the entire defense, the limiting factor for yards gained is the distance to the end zone. These filters remove roughly 1200 data points.

The distributions of some important data are included below. Running plays are called most frequently with either 10 or short yards-to-go, which goes hand-in-hand with the number of running plays strictly decreasing as down increases. Coaches will run the ball most often on 1st and 10, or when a first down is likely. Running plays are called most often around the 25 yard line or near mid-field. After a kickoff, if a return is not attempted, the ball is placed on the 25 yard line. When the ball is near the 50 yard line, coaches will use the entire playbook, as there are less situations where a run play will result in a safety, long yards-to-go situation, or a punt on the next play. Run plays are also called less as the season goes on, as teams are making final pushes for playoff eligibility. The passing game becomes more important. This trend swiftly reverses in the final week, as teams try to avoid injuries before the playoffs or off-season.

Figure 3: Down

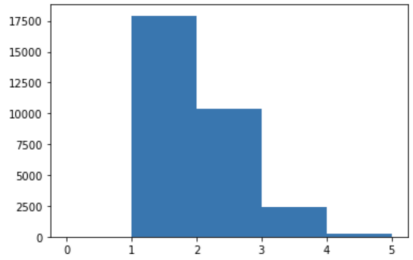


Figure 4: Yards-to-go

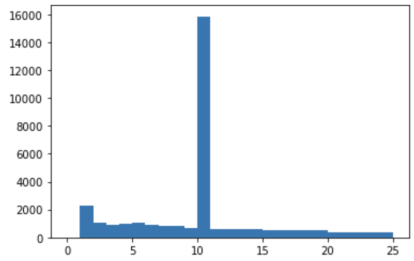


Figure 5: Player Height

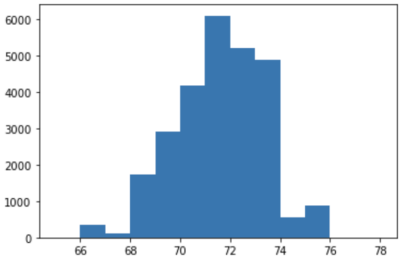


Figure 6: Player Weight

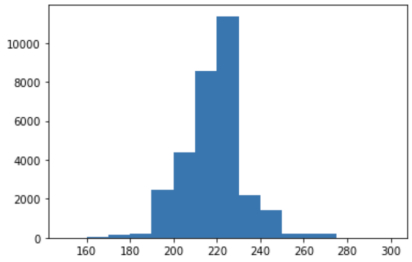


Figure 7: Yard line

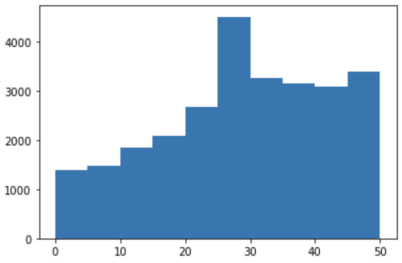
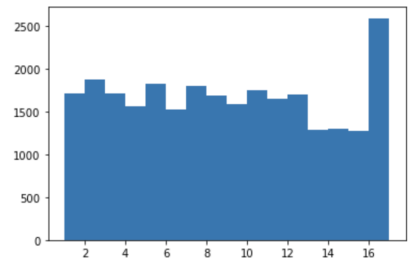


Figure 8: Week



3.2 Linear Regression

This section will contain two methodologies, one taking a frequentist approach to linear regression and the other taking a Bayesian approach. Advantages of the Bayesian approach was argued in Section 2.2 and will be reinforced in Section 4.2.

3.2.1 OLS Linear Regression

Ordinary least squares linear regression assumes data is generated from a model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, where $\mathbf{Y}$ is the response variable matrix, $\mathbf{X}$ is the "design" matrix, β is a matrix of weights, and $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is Gaussian noise. OLS makes the crucial assumptions that the model is linear with regression coefficients and noise, observations are uncorrelated with each other and the noise, noise is normally, identically, and independently distributed with mean 0 and constant variance, among others. OLS gives a closed form solution to $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, with predictions given by $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ This paper will use the `lm` method in R to construct an baseline OLS regression, run diagnostics, and select features.

3.2.2 Bayesian Linear Regression

Bayesian linear regression incorporates probability distributions in its analysis. Rather than using single points as the predicted outcome, it gives a probability distribution for the posterior probability of a predictor. As argued earlier in this paper, created a probability distribution would allow coaches to compare play calls and game plan based on probability of gaining a certain number of yards on each attempt.

Bayesian linear regression assumes that the coefficients have a prior multivariate normal distribution, as in, $\beta \sim \mathcal{N}(\beta_0, \Sigma_0)$, where $\beta_\mathbf{0}$ is a mean vector and $\mathbf{\Sigma_0}$ is the covariance matrix between the

coefficients. In this case, the data suggested the regressors were very uncorrelated, and we will assume that $Cov(\beta_i, \beta_j) = 0$ for $i \neq j$. We also specify priors for σ^2, the variance of the noise. I will use the half-normal distribution as the prior for σ^2.

Specifically, the final model is:

$$\begin{aligned}
\mathbf{Y} &\sim M\mathcal{N}(\mathbf{X}\beta,\, \sigma^2 I) && \text{Yards gained} \\
\beta_0 &\sim \mathcal{N}(0,\, 10) && \text{The Intercept} \\
\beta_1 &\sim \mathcal{N}(0,\, 10) && \text{Player Weight} \\
\beta_2 &\sim \mathcal{N}(0,\, 10) && \text{Down} \\
\beta_3 &\sim \mathcal{N}(0,\, 10) && \text{Yard Line} \\
\beta_4 &\sim \mathcal{N}(0,\, 10) && \text{Yards-to-go} \\
\beta_5 &\sim \mathcal{N}(0,\, 10) && \text{Temperature} \\
\sigma^2 &\sim HalfNormal(10)
\end{aligned}$$

Posterior distributions constructed through the "No-U-Turn Sampler" (NUTS). While in class we used Gibbs sampling to update posterior beliefs for β, σ^2, NUTS produces similar results to the Gibbs sampler but with faster computation times, less sensitivity to prior parameter choice, and gives larger effective sample sizes. Sampling will be implemented using the `PyMC` Python library. A discussion about why these variables were selected can be located in section 4.1. Very weakly informative priors were chosen for the coefficients. As there are 26,885 points in the data set, the data will dominate the likelihood, and using a nearly non-informative prior should still yield good results.

4 Data Analysis and Results

4.1 OLS Regression

A naive OLS Regression, using all numeric features, produced a poor fit, with a multiple R-squared 0.01403, and a mean squared error of 20.32. By plotting residuals, it can be seen that the homoscedasticity assumption of OLS may be violated, as variance appears to lower for both low and high predicted gains than medium gains. The model also showed that player height, week, humidity, and home and away scores were poor predictors of yards gained, as they all had p-values of > .1.

Humidity being a bad predictor makes sense, especially with the lack of information on precipitation. Coaches may call more run plays in the rain or when the ball is harder to catch due to pressure changes at low temperatures, but changing humidity with no precipitation should not affect run performance. Although number of rushing attempts was correlated with the week, performance in that week should not change, outside of factors like temperature and precipitation, of course. Oddly, points scored did not have much statistical significance in run performance. Football commentators emphasize "momentum", or more specifically, the mental aspect of a game. If a team is already performing poorly or well, they may continue to perform badly or succeed, respectively, because their feelings affect their performance. However, the data suggests this is not a big factor in yards gained.

Removing these statistically insignificant factors produced an R-squared identical to the original model but slightly decreased mean-squared error to 20.27.

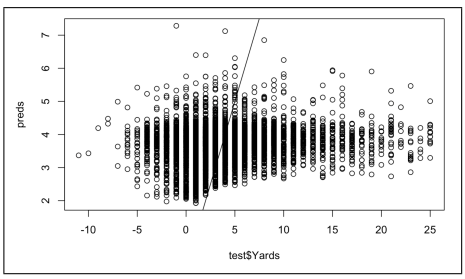


Figure 9: OLS Prediction vs. Fitted

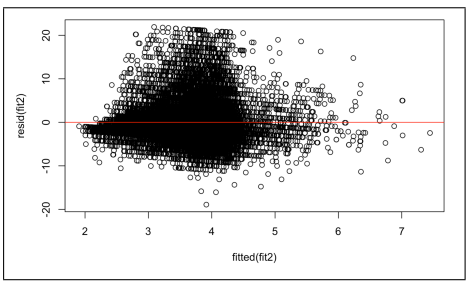


Figure 10: OLS Residual Analysis

The magnitude of these coefficients suggest that a heavier running back will perform slightly worse than a lighter one, high temperatures negatively affect performance, the closer to midfield the line of

scrimmage is more the more yards gained, and that the down and yards-to-go both have the largest impact.

4.2 Bayesian Linear Regression

Using the same five variables, I constructed a Bayesian linear regression model using the prior distributions listed in Section 3.2.2. The posterior prediction density for a run on the first play of the game after an opening kickoff touchback with fair weather can be seen below. That is, a player who weighs 200 pounds runs from the 25 yard line on 1st & 10, when the temperature is 70 degrees.

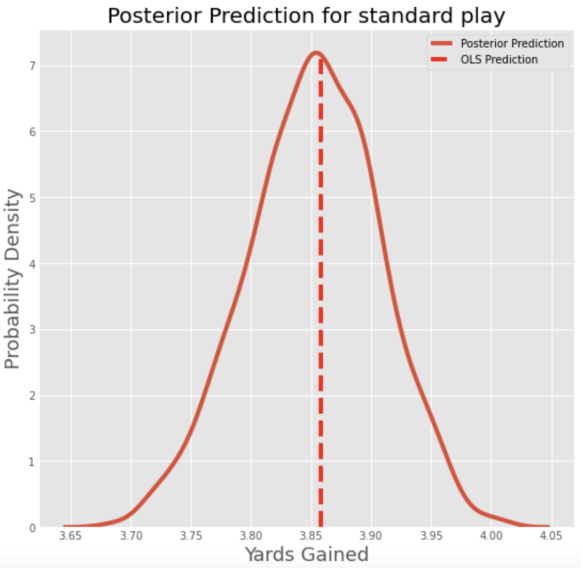


Figure 11: Posterior Prediction Distributions for a Standard First Play

OLS regression would predict he runs 3.86 yards. Austin Ekeler is the lead running back for the Los Angeles Chargers. He weighs roughly 200 pounds and plays home games in an indoor stadium, meaning he often has rushing attempts that exactly match this scenario. As of Week 15 of the 2022 NFL season, Ekeler is averaging 4.2 yards per attempt. This output suggests that a player with Ekeler's attributes would have a 0.3% chance of rushing for more than 4 yards on this play. If this play occured on the 50 yard line, however, the analysis suggests that there is a 99.65% chance that Ekeler would run for more than his season average. If it was 3rd & 4, Ekeler would have a 47.55% of getting a first down, when the OLS predictor suggests he would only run for 3.9 yards and fail to convert. Another informative example is 4th and 3 on the 2 yard line. The OLS model suggests that the player would run the ball for 2.8 yards. Using the Bayesian model, there is a 3.15% chance that Ekeler would score a touchdown. This scoring chance would not be a possibility in the OLS model.

The coefficient for the intercept in the OLS model was 3.36, suggesting that a player is essentially guaranteed over 3 yards per attempt when not considering their attributes or the state of the game. The Bayesian analysis gives a posterior centered at 3.5, with a high density interval of [2.5, 4.4]. This sense of randomness helps take into consideration that even with knowledge of the defense and the scenario, those 3 yards are never guaranteed. The coefficients for player weight and temperature were slightly negative, suggesting that the heavier a runner is, and the higher the temperature, the worse yards per attempt they will have. In the Bayesian model, the HDIs were [-0.011, -0.0023] and [-.0069, -.001], respectively. This conclusion was consistent across both models. Both down and yards-to-go had a positive coefficient in the frequentist model, and in the Bayesian model, their posterior distributions were centered at 0.16 and 0.12 with HDIs of [0.068, 0.24] and [0.11, 0.14], respectively. In both analyses, later downs and further distances for a first down both yield higher yards gained. Defenses are more likely to expect a run on early downs and in short yardage situations, and they expect a pass on later downs and in longer yardage, as football intuition says the probability of getting a first down is higher with a pass. By breaking this expectation, NFL coaches who run on long yards-to-go and on later downs can surprise the defense and get better gains than the standard

run situations. However, the model neglects the fact that 4th down is usually reserved for punting or attempting a field goal, and lining up in a non-special-teams formation would alert the defense to expect an offensive play.

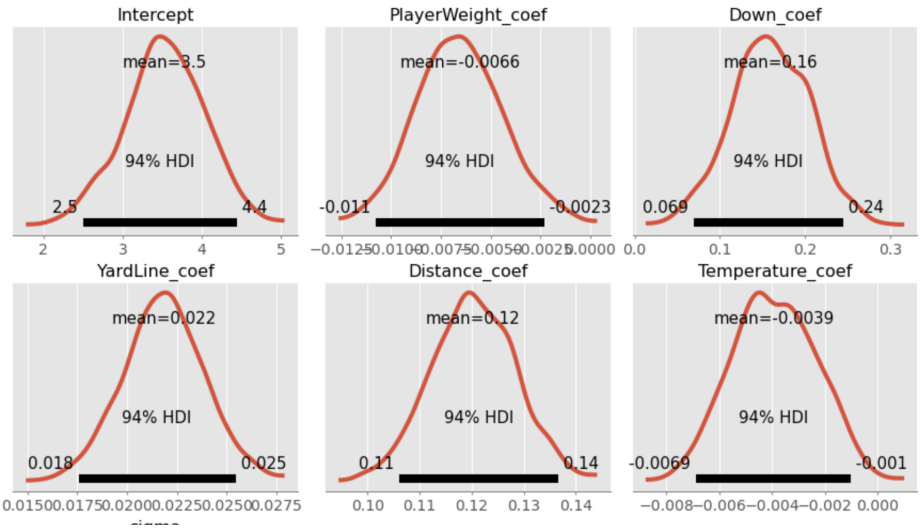


Figure 12: Posterior Distributions for Coefficients

The posterior density for σ can be seen below. Centered around 4.5, with the HDI being [4.475, 4.55], the posterior distribution suggests a high variance of the response variable. Further analysis should be done to reduce the variance, albeit to create a more biased model.

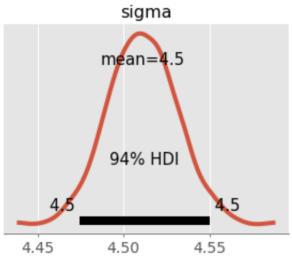


Figure 13: Posterior Distributions for variance

In the future, I hope to use Bayesian classification techniques, such as Bayesian logistic regression in the GLM family, to create posterior distributions for first downs and touchdowns. I also hope to include categorical data, such as offensive and defensive formation, to determine how pre-snap alignments affect posterior densities.

References

[1] Arkes, J. (2011). Is controlling the rushing or passing game the key to NFL victories? The Sport Journal, 14(1).

[2] Estimating the Causal Effect of Defensive Formation on Yards Gained in Run Plays. (n.d.). Retrieved December 22, 2022, from https://operations.nfl.com/media/4199/bdb_kruchten.pdf

[3] Gelman, A. (2020, April 17). Prior choice recommendations · Stan-Dev/Stan Wiki. Retrieved December 21, 2022, from https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

[4] Historical NFL rushing trends (and ERA adjustment issues). (2016, May 16). Retrieved December 21, 2022, from https://www.eldo.co/historical-nfl-rushing-trends-and-stat-adjustment-considerations.html

[5] Nishio, M., amp; Arakawa, A. (2019). Performance of Hamiltonian Monte Carlo and No-U-Turn Sampler for Estimating Genetic Parameters and Breeding Values [Abstract]. Genetics Selection Evolution, 51(1). doi:10.1186/s12711-019-0515-1

[6] Pelechrinis, K., amp; Papalexakis, E. (2016). The Anatomy of American Football: Evidence from 7 years of NFL Game Data. PLOS ONE, 11(12). doi:10.1371/journal.pone.0168716

[7] Tanier, M. (2016, November 16). N.F.L. Teams Have Rediscovered Their Running Backs This Season. New York Times.

[8] Addison Howard, Jay Evan Reid, Michael Lopez, Will Cukierski. (2019). NFL Big Data Bowl. Kaggle. https://kaggle.com/competitions/nfl-big-data-bowl-2020